DSSTox Log File:

Carcinogenic Potency Database Summary Tables – All Species (CPDBAS)

(last updated 10 April 2006)

Description: Information in this file documents the creation, review, and update process for the DSSTox CPDBAS SDF file, provides summary information on database contents, and lists currently unavailable CAS registry numbers for known structures. The first section summarizes the process used for creating the initial DSSTox SDF files and the quality assurance checks and procedures employed. A table providing field and data counts offers summary overview of CPDBAS file contents and chemical composition. A second table provides summary counts of various types of replicate chemical information in the CPDBAS file. The Log table documents any modifications and revisions to the database content or format in version updates. To obtain the most current version of this Log File and a record of any new modifications, or to report errors in this file, a user should consult the DSSTox CPDBAS database page: http://www.epa.gov/nheerl/dsstox/sdf_cpdbas.html.

QA and Development Notes for v1a:
CPDB SDF files underwent an extensive series of quality review checks prior to publication of initial launch versions. Source field entries (i.e. non-DSSTox Standard fields) were thoroughly checked by visual inspection for correspondence to original CPDB Summary Tables. We thank Lois Swirsky Gold and Thomas H. Slone for valuable assistance in ongoing quality review of the DSSTox CPDB files, helping to ensure that data are accurately extracted and represented from the original CPDB Summary Tables. They pointed out numerous systematic and human-error problems early in the DSSTox project and early in the process of CPDB SDF development, carefully reviewed DSSTox field definitions, offered suggestions for improving and finalizing all documentation files, and worked with the DSSTox team to find missing structures and reconcile remaining discrepancies in CAS numbers from the original CPDB Summary Tables.

Chemical structures were initially obtained by automated filling from large in-house databases of CAS-referenced structures (American Chemicals Directory, NCI Structure Database). The ChemFinder website (http://chemfinder.cambridgesoft.com/) was used extensively for checking CAS-to-structures and for retrieving CAS numbers for parent forms of salts and complexes. CambridgeSoft's ChemOffice 2002 ChemFinder (ver 7.0 for Windows) was used for automatic generation of SMILES codes from structures and both ChemFinder and ACD ChemFolder (ver 6.0 for Windows) were employed for "Structure-to-Name" or "Name-to-Structure" features. **ChemName**, **SMILES**, **CAS** and **Structure** field contents were checked by cross-referencing wherever possible. The CPDBRM_DOP_v1a (defined organic parent) SDF file was created by exporting only defined organics to SDF from the Main ChemFinder file for CPDBRM, and converting salts and complexes to their simplified form, with changes to corresponding Standard Chemical Fields. For versions 2 and later, a DOP is not created; rather the **SMILES_Parent** field is included and can be used to create a "desalted" version of the Main file by the user. All CAS registry numbers in the CPDBAS_v2a file were checked by the CAS check-digit verification algorithm (http://www.cas.org/EO/checkdig.html) using a Python script (CASlistcheckv2.py) created by Stephen Little (EPA).

Notes for v2a:
For version 2a, a variety of fields have been added. IUPAC systematic chemical names, **ChemName_IUPAC**, were computed by Marc Nicklaus (NCI) using the ACD Labs IUPAC Name-Generation software (ACD/NameBatch, version 8.05). Where IUPAC names were not provided, systematic names were either inferred from the structure or obtained from the TOXNET ChemID website (http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp). **INChI** codes were automatically generated from the final DSSTox SDF using a pre-release version of the publicly available program, wINChI11b.exe, accessible from the NIST INChI developers (http://chemdata.nist.gov/IChI/INChIv11b.zip). AuxInfo strings, which can be used to reproduce the molfile structure, are typically generated along with the INChI codes. However, due to their length frequently exceeding the 255 character limit of some

Chemical Relational Database applications and the non-unique nature of the AuxInfo text string, we include only the invariant INChI codes in the DSSTox data files.

In addition to the incorporation of two new DSSTox Standard Chemical Fields (**ChemName_IUPAC** and **INChI**), three new DSSTox Standard Toxicity Fields have been added to CPDBAS: **StudyType**=carcinogenicity; **Species**=rat, mouse, hamster, dog, rhesus, cynomolgus, bush baby, tree shrew; **Endpoint**=TD50, Tumor Target Sites. In the case of CPDBAS, the **Species** field enables listing of all species for which data are available for a given chemical. A number of additional modifications were incorporated into the CPDBAS_v2a to facilitate the use of these data in structure-activity studies and relational database searching. The 4 separate data files published in the original DSSTox CPDB v1a (CPDBRM, CPDBDG, CPDBHA, and CPDBPR) have been consolidated into a single file (CPDBAS = CPDB All Species) containing a total of 1451 records. Data in this file were extracted from the CPDB Summary Table files posted on the Source Website as of 15Nov04: 1433 records from the updated RatMouse Summary Table and an additional 18 unique chemical substances from the other 3 species tables that were not included already in the RatMouse table. A total of 9 records that were included in CPDB v1a (all starches and pectins) no longer appear in the current CPDB Summary Tables and so have been deleted from CPDBAS_v2a, whereas a total of 88 new records from the CPDB RatMouse Summary Table and 3 new records from the CPDB Hamster Summary Table have been incorporated into CPDBAS_v2a. CPDBAS_v2a also incorporates a number of data modifications in the TD50 and Target Sites fields for several chemical records that were included in v1a, these modifications obtained from the most current CPDB Summary Table files posted on the Source Website as of 15Nov04. Wherever new data were added or modifications were made to existing toxicity data from CPDB v1a, a notation is included in a new field, ToxNote. New data fields were created to accommodate the data from the four original data tables. The only exceptions are the TD50 and Target Sites fields for primate species "bush babies" and "tree shrews", which are not included as separate fields in v2a since they each have only a single data entry in 1451 records. Instead, these data are indicated in the **Species** field and listed in the ToxNote field entry for the corresponding chemical. CPDBAS_v2a also includes 3 new, purely numeric fields for the largest TD50 data columns, i.e. **TD50 Rat notext**, **TD50 Mouse notext**, **TD50 Hamster notext**, alongside the original field, the latter including text notes and the lettered footnote references from the original CPDB Summary Tables. The pure numeric fields are intended to facilitate exploration of numerical trends of the TD50 as a function of chemical structure. Finally, we have discontinued offering the DOP (defined organic parent) file containing simplified-to-parent structures for the defined organics. Instead, we offer a number of standard chemical fields, including **SMILES_Parent**, which should enable a user to create easily their own “desalted” parent file for specialized purposes.

Finally, in version 2a, a number of corrections and modifications of Standard Chemical Fields, including structures, have been made. These include the addition of 2D stereochemistry indicators in many records that were found to overlap with other DSSTox databases containing stereochemistry, in particular, NCTRER and FDAMDD. Due to the sheer number of modifications made to CPDBAS_v2a versus the earlier 1a versions, we do not list each of these corrections separately; although v2a record DSSTox_IDs are listed where there was an error in the v1a structure. It is recommended that users replace the earlier 1a versions with this new v2a in its entirety.

Notes for v3a,b:

CPDBAS_v3 has 31 new chemical records added from v2a and data were added to several existing records from v2a. Revised DSSTox Standard Chemical Fields are included (see http://www.epa.gov/nheerl/dsstox/MoreonStandardChemFields.html) along with updated InChI codes (version 1.0), recomputed IUPAC chemical names (ACDLabs ACD/Name, version 9.0), and many regenerated 2D structures with stereochemistry of steroidal compounds rendered in more standardized form. With the modifications to the DSSTox Standard Chemical Fields, to better distinguish **STRUCTURE** fields from **TestedForm…** fields, many more structures are included in CPDBAS for substances classified as mixtures, e.g. **STRUCTURE_Shown =** “active ingredient of mixtures”, “monomer of polymer”, “representative isomer of mixture”, representative component of mixture”.

Some CPDBAS Source Toxicity Fields were deleted and several new fields were added (see below). In addition, many field names were changed to eliminate spaces and provide more descriptive names. Finally, an extensive quality review of all DSSTox chemical records was performed, resulting in numerous corrections and modifications to chemical structures and added information (CASRN, representative structures for mixtures, etc) throughout DSSTox data files. For more information on current review procedures, see http://www.epa.gov/nheerl/dsstox/ChemicalInfQAProcedures.html

Log of SDF Modifications and Version/revision updates:

Date	DSSTox SDF File Name	Modifications from previous version	Additional Notes
15Oct03 15Oct03 15Oct03 15Oct03 15Oct03	CPDBRM_v1a_1354_15Oct03.sdf CPDBRM_DOP_v1a_1189_15Oct03.sdf CPDBHA_v1a_80_15Oct03.sdf CPDBDG_v1a_5_15Oct03.sdf CPDBPR_v1a_27_15Oct03.sdf	Initial launch publication; no previous versions.	Working with Source collaborators (L.S. Gold and T. H. Slone), periodic version updates to the DSSTox CPDB SDF files (i.e., v1, v2, etc.) will incorporate new information provided in updates to the CPDB Summary Tables and posted on the Source CPDB website, http://potency.berkeley.edu/. In addition, revision updates (e.g., v1a, v1b, etc) will correct reported errors or add missing data provided by users or the Source.
29Mar04	CPDBPR_v1b_27_15Oct03.sdf	Corrected structure DSSTox_ID=2, 2,7-Acetylaminofluorene	Thanks to ACD Labs
1Mar05	CPDBAS_v2a_1451_1Mar05.sdf	Consolidation of data from 4 previous v1 files (CPDBRM, CPDBHA, CPDBDG, CPDBPR) into single CPDBAS_v2 file; Addition of 3 pure numeric fields, **TD50 Rat notext**, **TD50 Mouse notext**, **TD50 Hamster notext**, for specialized use. Replacement of field entry "ND" (no data) with blank entries in data fields. Addition of 88 new chemical records for Rat or Mouse and 3 new chemical records for Hamster. Also, modification of data fields for many previous v1 chemical records. Modifications extracted from the current CPDB Source Website Summary Tables (15Nov04). Addition of **SMILES_Parent** to Main file. New Standard Chemical Fields: **INChI**, **ChemName_IUPAC** New Standard Toxicity Fields: **StudyType**, **Species**, **Endpoint** Modified Field Names: **Target Sites Rat Both** **Target Sites Mouse Both**, **Target Sites Hamster Both**,	Major format modification to include INChI, IUPAC names, and ToxML fields. Separate "desalted" defined organic parent (DOP) file not provided. Users can easily generate DOP file by extracting "defined organic" records and converting **SMILES_Parent** to structures. Since both tree shrew and bush baby (Non-Human Primates) each had only a single record for which data were available, and these records already contained data for other species, separate fields for these species were not included; rather these data are noted by the entries "bush baby" or "tree shrew" in the **Species** field and TD50 and Target Sites are listed in the **ToxNote** field of the corresponding records.

		to **Target Sites Rat Both Sexes**, **Target Sites Mouse Both Sexes**, **Target Sites Hamster Both Sexes**. Deleted Fields: **TD50 Tree Shrews** and **Target Sites Tree Shrews**, **TD50 Bush Babies** and **Target Sites Bush Babies** **OtherSpecies** Additional Toxicity Field: **ToxNote** Corrected errors in structures for CPDBAS_v2a:DSSTox_ID = 82, 120, 565, 603, 605, 663, 670, 798, 821, 880, 1050, 1144	
10Apr2006	CPDBAS_v3b_1481_10Apri2006.sdf	Updated with new DSSTox Standard Chemical Fields and entries (*revised Aug 2005*). Updated InChI codes (version 1.0). Updated IUPAC chemical names (ACDLabs Name to Structure, version 8.0). Expanded "ddmmmyear" format for dates in DSSTox file names (e.g., 10Apr2006). Addition of 30 new chemical records. Also, modification of data fields for many previous v2 chemical records. Modifications extracted from the current CPDB Source Website Summary Tables (01Jun2005). Deleted Source-related fields: **TD50_Rat_notext** **TD50_Mouse_notext** **TD50_Hamster_notext** New Source-related fields: **TD50_Rat_mmol** **TD50_Mouse_mmol** **TD50_Hamster_mmol** **ActivityCategory_SingleCellCall** **ActivityCategory_MultiCellCall** **NTP_TechnicalReport** **Website_url** Renamed Source-related fields to be more descriptive and eliminate spaces:	Numerous structure modifications and changes in stereochemical rendering throughout DSSTox data files following major quality review. CPDBAS_v3a _1484_22Oct2005: ***Note:*** Earlier version of this file was provided to PubChem, with identical format to v3b but latter has undergone additional QA review and has a small number of corrections/modifications. Also, 3 duplicated records were reconciled and eliminated in subsequent version.

		Mutagenicity_SAL_CPDB **TD50_Rat_mg** **TargetSites_Rat_Male** **TargetSites_Rat_Female** **TargetSites_Rat_BothSexes** **TD50_Mouse_mg** **TargetSites_Mouse_Male** **TargetSites_Mouse_Female** **TargetSites_Mouse_BothSexes** **TD50_Hamster_mg** **TargetSites_Hamster_Male** **TargetSites_Hamster_Female** **TargetSite_Hamster_BothSexes** **TD50_Dog_mg** **TargetSites_Dog** **TD50_Rhesus_mg** **TargetSites_Rhesus** **TD50_Cynomolgus_mg** **TargetSites_Cynomolgus** **ToxicityNote**	

Field and Data Counts in DSSTox SDF files: Refer to CPDBAS_FieldDefFile for definitions and explanations of all terms.

DSSTox SDF	Standard Chemical Fields	Standard Toxicity Fields	Source-specific fields	Chemical records total	Defined organic	Inorganic	Organo-metallic	Mixture or unknown*	Parent	Salt or Salt complex	Complex
CPDBRM_v1a	14	0	10	1354	1189	52	39	74	1016	99	165
CPDBRM_DOP_v1a	16	0	10	1189	1189	0	0	0	1000	67	122
CPDBHA_v1a	13	0	6	80	72	6	1	1	67	5	7
CPDBDG_v1a	13	0	4	5	5	0	0	0	4	1	0
CPDBPR_v1a	13	0	10	27	24	1	0	2	21	3	1
CPDBAS_v2a	17	3	23	1451	1280	56	40	75	1089	102	194

* All substances classified as **SubstanceType** = “mixture or unknown” in the CPDB data files are definitively known to be mixtures or formulations; there are no unknowns.

CPDBAS SDF Content*	Totals_v3b
# Records	1481
DSSTox Standard Chemical Fields	18
DSSTox Standard Toxicity Fields	3
CPDBAS Source Fields	27
Total # Fields	48
Chemical Content	**Counts_v3b**
STRUCTURE_ChemicalType:	
defined organic	1344
inorganic	58
organometallic	42
no structure	37
STRUCTURE_TestedForm_DefinedOrganic:	
parent	1122
complex	151
salt	74
salt complex	3
TestSubstance_Description:	
single chemical compound	1385
defined mixture or formulation	58
undefined mixture	27
macromolecule	10
unspecified or multiple forms	0

Replicate Information in CPDBAS_v3 SDF File: The term “replicate” refers to possibly redundant information in the chemical structure fields. All replicate cases can be easily located by search of the **ChemicalNote** and **ChemicalReplicateCount** fields in CPDBAS (refer also to CPDBAS_FieldDefFile).

CPDBAS: Replicate Type	Sets of Replicates	Individual Cases
CAS [1]	13	29
2D structures [2]	8	18
Parent structures [3]	30	61
Totals	51	108

[1] replicate CAS: same CAS number (e.g., if different technical grades or related mixture formulations were tested for carcinogenicity).

[2] replicate 2D structure: geometric or stereoisomers (e.g., cis and trans, RS, dl forms that might provide the same 2D information)

[3] replicate parent structures: salt or complex of same parent structure (e.g., Na and K salt of same parent structure)

Wanted!! CASRN Information

The listing below provides chemicals with known structures and Unknown CASRN entries, which is primarily an indication of the little studied nature of these particular chemicals in the CPDB. For each, a CAS registry search was performed in CAS SciFinder and no CASRN was found by the CPDB Source authors. However, if a user has new information pertaining to any Unknown CASRN in the below listing, please report this using a DSSTox Error Report Form that can be accessed from any DSSTox SDF Download Page, and be sure to indicate all relevant information (full DSSTox SDF file name, **DSSTox_ID_FileName**, **TestSubstance_ChemicalName**, nature of missing information, source of correct information, etc.). Thank you!

TestSubstance_ChemicalName	STRUCTURE_SMILES	CASRN	Date of Request
3-Amino-4-[2-[(2-guanidinothiazol-4-yl)methylthio], ethylamino]-1,2,5-thiadiazole	N=C(N)NC1=NC(CSCCNC2=NSN=C2N)=CS1	Unknown	10Apr2006
2-Azoxypropane	[N+](=N\C(C)C)(/C(C)C)[O-]	Unknown	10Apr2006
1-Chloroethylnitroso-3-(2-hydroxypropyl) urea	N(C(=O)NCC(C)O)(N=O)C(C)Cl	Unknown	10Apr2006
3-Diazotyramine.HCl	C1(/C=C(\C=C/C1=O)CCN)=[N+]=[N-].[HCl]	Unknown	10Apr2006
Diethylacetylurea	O=C(N(CC)CC)NC(C)=O	Unknown	10Apr2006
Dimethylaminoethylnitrosoethylurea, nitrite salt	N(C(=O)[NH3+])(CCN=O)CCN(C)C.N(=O)[O-]	Unknown	10Apr2006
N,N-Dipropyl-4-(4'-[pyridyl-1'-oxide]azo)aniline	N(=NC1=CC=C(C=C1)N(CCC)CCC)C2=CC=[N+](C=C2)[O-]	Unknown	10Apr2006
3-O-Dodecylcarbomethylascorbic acid	O(C1[C@@H]([C@@H](O)CO)OC(=O)C(O)=1)C(C(O)=O)CCCCCCCCCCCC	Unknown	10Apr2006
1-Ethylnitroso-3-(2-hydroxyethyl)-urea	O=C(N(CC)N=O)NCCO	Unknown	10Apr2006
1-Ethylnitroso-3-(2-oxopropyl)-urea	O=C(N(CC)N=O)NCC(=O)C	Unknown	10Apr2006

N2-gamma-Glutamyl-p-hydrazinobenzoic acid	N(NC(CC[C@H](N)C(=O)O)=O)C1C=CC(=CC=1)C(=O)O	Unknown	10Apr2006
Hexanal methylformylhydrazone	CCCCC/C=N/N(C=O)C	Unknown	10Apr2006
1-(2-Hydroxyethyl)-nitroso-3-ethylurea	O=C(N(CCO)N=O)NCC	Unknown	10Apr2006
IQ.HCl	CN1(C2=C(C3=C(C=C2)N=CC=C3)N=C1N).[H]Cl	Unknown	10Apr2006
MeA-alpha-C acetate	C1(=CC=CC2=C1NC3=C2C=C(C(=C3)[N+])C).CC([O-])=O	Unknown	10Apr2006
1-Methyl-1,4-dihydro-7-[2-(5-nitrofuryl)vinyl]-4-oxo-1,8-naphthyridine-3-carboxylate, potassium salt	[Na+].C1(=CC=C2C(=N1)N(C=C(C2=O)C([O-])=O) C)/C=C/C3=CC=C(O3)[N+]([O-])=O	Unknown	10Apr2006
3-Methylbutanal methylformylhydrazone	CC(C/C=N/N(C=O)C)C	Unknown	10Apr2006
Methylnitrosamino-N,N-dimethylethylamine	N(CCN(C)C)(C)N=O	Unknown	10Apr2006
(N-6)-(Methylnitroso)adenine	C12=C(NC=N1)N=CN=C2NCN=O	Unknown	10Apr2006
N,N-Dipropyl-4-(4'-[pyridyl-1'-oxide]azo)aniline	N(=NC1=CC=C(C=C1)N(CCC)CCC)C2=CC=[N+](C=C2)[O-]	Unknown	10Apr2006
N2-gamma-Glutamyl-p-hydrazinobenzoic acid	N(NC(CC[C@H](N)C(=O)O)=O)C1C=CC(=CC=1)C(=O)O	Unknown	10Apr2006
N-Nitrosomethyl-(2-tosyloxyethyl) amine	NC(CN=O)COS(=O)(C1=CC=C(C)C=C1)=O	Unknown	10Apr2006
N-Nitroso-ethylhydroxyethylurea	N(C(=O)N)(CCN=O)CCO	Unknown	10Apr2006
N-Nitroso-ethyl-2-oxopropylurea	N(C(=O)N)(CCN=O)CC(C)=O	Unknown	10Apr2006
N-Nitroso-oxopropylurea	N(C(=O)N)CCC(=O)N=O	Unknown	10Apr2006
N-Nitroso-oxopropylchloroethylurea	N(C(=O)N)(CCC(=O)N=O)CCCl	Unknown	10Apr2006
N-Nitroso-2-phenylethylurea	O=C(N(CCC1=CC=CC=C1)N=O)N	Unknown	10Apr2006
2-Oxopropylnitrosourea	N(C(=O)N)(N=O)CC(C)=O	Unknown	10Apr2006
Palonidipine.HCl	O=C(C1=C(C)NC(C)=C(C(OCC(C)(C)CN(CC3=CC=CC=C3)C)=O)C1C2=CC([N+]([O-])=O)=CC=C2F)OC.HCl	Unknown	10Apr2006
PhIP.HCl	N1(=C2C(=CC(=C1)C3=CC=CC=C3)N(C(=N2)N)C).[H]Cl	Unknown	10Apr2006